DOCUMENT RESUME

ED 049 273 | TM 000 445

AUTHOR	Egelston, Richard L.
TITLE	Test Achievement: Expectation and Reality.
PUB DATE	Feb 71
NOTE	21p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 1971
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Academic Performance, *Achievement, *Aspiration, *Curriculum Development, Educational Needs, Grade Prediction, Performance Criteria, *Prediction, Reinforcement, Sciences, *Self Evaluation

ABSTRACT

Levels of aspiration and student predictions as applied to test performance were employed in this longitudinal investigation of the process of self-evaluation. Two hundred and ten students from a rural secondary school in general and earth science classes were grouped according to previously demonstrated academic ability. Throughout the school year, the students were asked to predict the percentage score they would receive on each unit test they took immediately before and after its administration. Although explicit instructions about how to make predictions were not given, several students were able to improve their predictions over time. More able students tended to be more accurate in their predictions than the less able; and there appeared to be no sex differences operating. Trend analyses were conducted to ascertain the effect of practice upon learning how to make realistic predictions. The rate of improvement tended to be higher for high ability students, who gained the most from repeated performance. It is suggested that, since the study was limited to the familiar task of test taking, students were more likely to assess their performance accurately on this activity than on a less familiar one. Because many important decisions must be made by the individual, on the basis of ability and interests, after he has left the formal educational setting, a strong recommendation is made for the teaching of self-appraisal techniques within the regular school curriculum. The science classes are proposed as a logical place to start such instruction. (TA)

Test Achievement : Expectation and Reality

by

Richard L. Egelston
Psychology Department
State University College
Geneseo, N. Y. 14454

A paper presented at the 1971 Annual Meeting of the American Educational Research Association held in New York City, February 4-7, 1971; session C8 Learning in the Classroom.

When people leave the formal educational setting and enter the worlds of work and leisure, they are required to make many decisions based upon their own abilities and interests. Each of the decisions requires some assessment about the degree of success or enjoyment in the activity in which they are to become engaged. Hopefully, the evaluation of the potential activity will be rational and based upon a thorough knowledge of personal capabilities. However, experience indicates that self-evaluation is as difficult to learn as any other concept, and perhaps self-appraisal techniques need to be developed and taught within the school curriculum.

Research on self-evaluation is meager, and that which has been done generally involves simple tasks not at all comparable to the complex activities that individuals later undertake. Furthermore, few studies of a longitudinal nature have been undertaken.

The technique for studying level of aspiration was developed by Lewin and his students (Rotter, 1942) and involves a variable called a discrepancy score. The discrepancy score is defined as a difference between some expected or predicted score and some achieved score. Some researchers use a discrepancy between achievement on event A and predicted achievement of event B. Others use the discrepancy between achievement on event A and predicted achievement of event A. This technique may also be used to study self-evaluation.

Some important determinants of level of aspiration are brought out by Lewin (1936). According to Lewin, level of aspiration may be dtermined by the upper limit of the person's achievements (ability) and by the level of achievement of his social group (peer group). A third determinant may be the relative success of the individual in accomplishing similar goals in the past.

Murstein (1965) found that neither high nor low achieving college students changed their predictions of final grades as a result of midsemester performance. This result was contradicted by Wolfe (in press) who found that college students became more accurate predictors as a result of midsemester feedback.

Pennington's (1940) experiments on college students indicated that failure resulted in a lower level of aspiration, and success (passing with high grades) resulted in an upward swing in predicted scores on the following examination. With fifth grade children, Anderson and Brandt (1939) found that poor students set goals consistently above past performance, and good students set goals consistently below past performance.

In an attempt to determine the influence of sex and achievement on the ability to predict test scores for college students, Sumner and Johnson (1949) found discrepancy scores to be less for high achieving students than for low achieving students. They also found that females of all quartile levels are more accurate predictors than males of a comparable level.

With secondary school students Pickup and Anthony (1968) found that females who predicted higher scores than they received tended to reduce subsequent predictions while males did not. Low achievers were more likely to predict higher scores than received than high achievers.

Classroom measurements from test predictions may suffer from the experimenter effect. Research completed by Rosenfeld and Zander (1961) indicate that the level of aspiration may be influenced by reward or power. The rewards may be given via non-verbal cues emitted by the teacher in advance of and/or during the testing situation.

Method

Two hundred ten students in eight general science classes and one earth science class from a rural Eastern New York secondary school were used as subjects. Classes varied in size from sixteen to thirty-two students and were taught by two teachers. Within each grade students were grouped by academic ability from previous performance. The top one-fourth of the students in each grade were grouped for enrichment courses and the remaining students were divided into two sections of comparable ability.

At the beginning of the school year the teachers explained to the students that on each unit test the students would be asked to predict the percentage score they would get on the test immediately before and immediately after taking it. Separate slips of paper were stapled to the test for the pretest guess, and when completed were torn off and collected. Space was available on the test booklet for recording the post-test predictions. Both predicted scores and the actual scores were transferred to permanent record sheets. Since precentage grades were used district-wide as the method of reporting academic progress, the format for making predictions was not unfamiliar to the students. The random variable employed was a discrepancy score which was defined as the absolute difference between a predicted score and the obtained score.

The number of tests given to each class ranged between eight and thirteen. All tests were constructed to be discriminatory in nature, and perfect scores were rarely achieved. Thus, ceiling effects were not a contaminating variable. However, report card grades were adjusted to account for the test difficulty.

Students were told to base their predictions upon how well they understood the material and how difficult they thought the test would be (or was). Reminders were frequently given that the predictions would not affect actual grades in any way.

In the few cases where the subject failed to make a prediction, the mean prediction was used and was derived from all the pretest or posttest predicted scores the subject did make.

Within each section subjects were ranked from high to low on the final examination. Each section was then divided into four parts called quartiles. Within each section, however, the quartiles contained unequal n due to tied scores and the total section size not being divisible by four. Thus, the trend analyses were non-orthogonal. In only one section was the ratio of largest to smallest n as large as two.

One of the uncontrollable variables may have influenced the predictions at the beginning of the year. During previous years of schooling the students may have been accustomed to grades ranging from a low of sixty to one-hundred percent (failure set at seventy-five). Since the effective passing grade had suddenly been shifted from seventy-five to fifty percent by the teachers in the experiment, the grades achieved were lower in most cases. This unfamiliar situation may have caused the predicted grades to be much higher at the beginning of the year than they were at the end. No analysis of this variable was attempted.

RESULTS

Within each section a two-way factorial ANOVA was conducted using the quartiles as one main effect and the time of prediction (pretest and posttest) as the other. The data were pooled across all tests for each section. Table 1 presents the findings of the nine ANOVA'S with the significance level set at .05. Table 2 presents the ANOVA for section 9.

Insert Tables 1 and 2 about here

Significant differences were found among the quartiles within seven of the nine sections and between the two times of prediction for three sections. No significant interactions were found.

It might be concluded that even considering several trials individual differences will be maintained and that some students will be able to predict scores more accurately after completing the task after several practice trials than other students. Generally speaking, having completed the task will not allow for a more accurate self appraisal (before feedback) than prior to the task. Furthermore the relative improvement from pretest to posttest prediction remains relatively constant for all ability students.

In order to more completely examine the effect of practice upon learning how to make realistic predictions, trend analyses were conducted within each section. The assumption was made that each practice trial resulted in an equal amount of learning. The trend analyses were conducted upon the first and third levels of a three way factorial design: quartiles (A) by time of prediction (B) by test number (C). The analyses were complicated by the fact that the quartiles were of unequal size requiring a non-orthogonal analysis technique. In each case the hypotheses were tested in the following order: cubic interaction (AXC), cubic trend (C), quadratic interaction (AXC), quadratic trend (C), linear interaction (AXC), linear trend (C), and the contrast of the first and last predictions (C). Unfortunately with non-orthogonal analyses, the order of hypothesis tested is important since the tests of significance are not independent. In the four cases of multiple significant findings the set of hypothesis tests was not reordered to verify subsequent results. Note, however, that the tests of the first two main effects had been conducted prior to the trend analyses. Residual terms were not tested for significant higher order effects. Table 3 summarizes the trend analyses for the nine sections. At least one significant trend compon-

---

Insert Table 3 about here

---

ent or contrast was found for eight of the nine sections.

Since the unit examinations were of differential difficulty, it was reasonable to expect that a simple function would not be found to describe the trend when the design was collapsed on the A effect (all subjects) and on the B effect (both pretest and posttest predictions). The expectation was borne out when at least a third degree polynomial was needed to describe the trend for four sections, and with four other sections a polynomial of at least the fourth degree would be needed.

Although previous analyses (see Table 1) indicated that seven of the nine sections had differences among the quartiles (pooled across tests) only four indicated differences on the interaction components of the trend analysis which were tested. The apparent contradiction may be explained by the higher order components of the trend interaction which were not tested. (For example, in section 7 there are four levels of A and 8 levels of C making 21 degrees of freedom for the interaction term. Only the linear component of the A effect was combined with the linear, quadratic, and cubic components of the C effect. The higher order effects would be at least quadratic in A and quartic in C simultaneously.) It might be concluded that the students in the quartile levels learn at varying rates and that these differences can be described by a linear function in less than one-half of the sections.

Within the same trend analyses, contrasts of the last predictions with the first predictions were conducted, and found to be more accurate at the end of the year in seven of nine sections.

Thus, with practice and without instruction as to "how" most students were able to improve their ability to evaluate their own performance. The distribution of discrepancy scores for each time of prediction (pooled across all tests and all sections) is given in Table 4.

Insert Table 4 about here

Two way analyses of variance (sex by time of prediction) were performed after pooling data across tests and quartiles within each section. In no instance was a significant difference found between males and females.

Of the 210 students 24 made pretest predictions within 5 points of their actual score at least one-half of the time. On the posttest predictions the number increased to 42. At the other end of the spectrum 64 students were off by at least 15 points one-half (or more) of the time on the pretest predictions. This number decreased to 37 on the posttest predictions. When the four frequencies are placed in a table (see Table 5a), the resulting chi-square value for a test of ...endence (11.64) was significant at the .05 level. This apparently contradictory result stems from the fact that the chi-square analysis was based upon data pooled across all sections while the analyses of variance were done on each section independently (three of the nine analyses were significant; see Table 1).

When the pretest and posttest frequencies (see Table 5b) of those within 5 points were divided into high and low achievers (within their section) and again analyzed with a test for independence, the chi-square value of 5.50 was significant at the .05 level. A similar analysis on the other set of frequencies (see Table 5c) failed to yield a significant chi-square value. Thus, it may be concluded that some students will profit from experience while others will not, but the more able students have a higher likelihood of improvement.

Insert Tables 5a, 5b, and 5c about here

According to Rotter (1942) and others, predicted scores are often dependent upon the actual performance of the previous trial. However, in the situations for which they postulate this score, the task from trial to trial is identical. In the present experiment the predictions are based upon new cognitive understandings for each trial. Since achievement scores are somewhat related from test to test, it is not unreasonable that predictions will be related to one another, and that discrepancy scores will be mediated by both achievement and previous predictions. The assumption was made that discrepancy score for trial t+1 was conditional upon the discrepancy score for trial t.

A vector of discrepancy scores was constructed for each student and the data coded as conditional frequencies with five point intervals. The data for all students in each section were pooled and conditional probability matrices (transition matrices) were derived. A Markov chain analysis provided limiting vectors of probabilities (tolerance = .0005) for each section. (The limiting vector provides an estimate of the proportion of time the group will predict any category over an infinite number of trials.) The limiting vectors were converted to cumulative probability vectors and the pretest vector was compared with the posttest vector with a Kolmogorov-Smirnov Two Sample Test. Table 6 illustrates the cumulative proportion vectors for each section and for all sections combined. Only sections 7, 8, and 9 and the combined group produced vectors which were significantly different at

Insert Table 6 about here

the .05 level. Table 7 illustrates the transition matrices for the pretest and posttest predictions for the combined groups. In each case where significance was found the cumu-

Insert Table 7 about here

lative probabilities in the lower categories was larger for the posttest prediction which indicates students (at least in grade 9) learn at a faster rate following the task than they do prior to the task. Perhaps a certain level of maturity is required for self-evaluation accuracy.

CONCLUSIONS AND IMPLICATIONS

The ability to accomplish accurate self-evaluation appears to be a rarely encountered phenomenon in the junior high school, but there is some evidence that students of this level can learn how to do it. In this experiment explicit instructions about how to make predictions were not given, but several students were able to improve their predictions over time anyway. Although there were no differences by sex, the more able students tended to be more accurate than the less able students. Furthermore the rate of improvement tended to be faster for high ability students. However, being a high ability student in no way guarantees his being able to discover how to accuragely assess his performance, and being a low ability student does not insure his being unable to discover the process. As might be expected, those students who were relatively accurate at the start of the experiment tended to gain the most from the repeated practice.

Evaluation following performance tends to be more accurate than evaluation prior to performance, but the evidence is not clear about this point. Inhelder and Piaget (1958) have produced inconclusive evidence that concepts are more effectively used by adolescents than by younger people. They found that formal reasoning begins to appear about age 11 or 12, and builds up to a plateau at age 14 or 15. Evidence from the present study seems to support Inhelder and Piaget, but the age groupings within each grade were not as clear as they should have been to fully examine their contention.

This study was limited to the familiar task of test-taking. Since all students have taken many tests it is reasonable to conclude that students are more likely to assess their performance accurately on these activities than on those with which they are unfamiliar.

Although we know t... transfer of training rarely takes place unless it is taught as a separate technique, this author believes that self-evaluation techniques are not taught in any form in the educational program today. With the great emphasis today on objective decision making, it would seem important to examine personal capabilities and personal performance in an objective light. It would also appear that the science classes would be the logical place to undertake instruction on self-evaluation since objective measurement forms one of the cornerstones of this field.

TABLE 1

Results of the Nine Analyses of Variance
Quartiles by Time of Prediction

Section	Quartile	Time	Interaction
1	<.05	ns	ns
2	<.05	ns	ns
3	ns	ns	ns
4	<.05	ns	ns
5	<.05	<.05	ns
6	ns	ns	ns
7	<.05	<.05	ns
8	<.05	ns	ns
9	<.05	<.05	ns

TABLE 2

Analysis of Variance for Section Nine Quartiles by Time of Prediction

Source	SS	df	MS	F	p
Quartiles	2453.04	3	817.68	9.69	<.05
Time	1389.89	1	1389.89	16.46	<.05
Interaction	226.29	3	75.43	.89	ns
Within Cell	55899.59	662	84.44		
Total	59968.81	669			

TABLE 3

Summary of Trend Analyses (non-orthogonal) for Each Section*

Section	N Tests	Q_1 n_1	Q_2 n_2	Q_3 n_3	Q_4 n_4	Cub Int	Cub	Quad Int	Quad	Lin Int	Lin	1st vs last contrast
1	8	5	4	5	5	ns	<.05	ns	<.05	ns	ns	<.05
2	10	4	5	6	3	ns	<.05	<.05	<.05	ns	<.05	<.05
3	9	5	4	3	4	ns	ns	ns	ns	ns	ns	<.05
4	8	6	6	7	5	ns	ns	ns	ns	<.05	ns	<.05
5	8	7	8	6	7	ns	ns	ns	ns	ns	ns	ns
6	11	4	4	5	4	ns	ns	ns	ns	ns	<.05	ns
7	8	8	7	7	8	ns	ns	ns	ns	<.05	ns	<.05
8	12	8	8	10	6	ns	.05	ns	ns	<.05	<.05	<.05
9	13	6	7	7	6	ns	.05	ns	ns	ns	<.05	<.05

*Linear, Quadratic, and Cubic tests are on the third (tests) effect, and the three interactions are on the quartile by tests factors.

TABLE 4

Frequency Tabulation of Discrepancy Scores
For Each Time of Prediction

	Discrepancy Interval					
	0-5	6-10	11-15	16-20	21-25	> 25
Pretest	556	434	270	254	167	300
Posttest	643	453	310	227	148	220

TABLE 5a

Frequency Table for Good and Poor Predictors

	Good ($\leq$ 5)	Poor ($\geq$ 16)	
Pretest	24	64	$\chi^2 = 11.64$
Posttest	42	37	$p \leq .05$

TABLE 5b

Frequency Table by Achievement Level for Good Predictors

	Pretest	Posttest	
Top half	10	31	$\chi^2 = 5.50$
Bottom half	14	11	$p \leq .05$

TABLE 5c

Frequency Table by Achievement Level for Poor Predictors

	Pretest	Posttest	
Top half	19	9	$\chi^2 = .34$
Bottom half	45	28	p is ns

TABLE 6

Cumulative Proportion Vectors For Kolmogorov-Smirnov Two Sample Tests By Section

Section	0- 5	6-10	11-15	16-20	21-25	over 25
1 Pretest	.27	.43	.55	.68	.79	1.00
Posttest	.25	.46	.60	.72	.80	1.00
2 Pretest	.36	.64	.72	.85	.96	1.00
Posttest	.35	.61	.77	.85	.93	1.00
3 Pretest	.31	.59	.84	.92	.96	1.00
Posttest	.42	.69	.83	.92	.96	1.00
4 Pretest	.23	.48	.61	.74	.80	1.00
Posttest	.27	.50	.65	.75	.85	1.00
5 Pretest	.26	.47	.63	.73	.83	1.00
Posttest	.32	.55	.72	.80	.90	1.00
6 Pretest	.26	.45	.62	.74	.82	1.00
Posttest	.26	.46	.60	.74	.82	1.00
7 Pretest	.22	.40	.55	.68	.79	1.00
Posttest	.34	.56	.72	.85	.93	1.00
8 Pretest	.26	.42	.54	.74	.83	1.00
Posttest	.29	.52	.66	.80	.86	1.00
9 Pretest	.34	.61	.75	.85	.90	1.00
Posttest	.44	.67	.83	.90	.95	1.00
Combined						
Pretest	.28	.49	.64	.76	.85	1.00
Posttest	.33	.55	.71	.82	.89	1.00

TABLE 7

TRANSITION MATRICES AND LIMITING VECTORS FOR ALL SUBJECTS COMBINED ON PRETEST AND POSTTEST PREDICTIONS*

transition matrix for pretest predictions

	0- 5	6-10	11-15	16-20	20-25	over 25
0- 5	.307	.247	.166	.135	.058	.087
6-10	.274	.242	.158	.111	.071	.144
11-15	.322	.222	.117	.097	.117	.125
16-20	.243	.270	.134	.122	.074	.157
21-25	.226	.129	.129	.181	.090	.245
> 25	.248	.118	.118	.150	.118	.248

transition matrix for posttest predictions

	0- 5	6-10	11-15	16-20	21-25	over 25
0- 5	376	.252	.140	.098	.047	.087
6-10	.336	.214	.181	.101	.085	.083
11-15	.334	.264	.156	.090	.073	.083
16-20	.288	.255	.160	.099	.080	.118
21-25	.296	.193	.126	.178	.081	.126
> 25	.213	.127	.145	.154	.113	.248

limiting vectors for both predictions

	0- 5	6-10	11-15	16-20	21-25	over 25
Pre	.278	.216	.142	.128	.083	.151
Post	.327	.227	.153	.109	.072	.109

* N is 1788 and 1794 for pretest and posttest predictions respectively.

REFERENCES

.derson, H. H. and H. F. Brandt, "A Study of Motivation, Involving Self-announced Goals of Fifth-Grade Children and the Concept of Level of Aspiration, "Journal of Social Psychology, X (1939), pp. 209-232.

.helder, B. and J. Piaget, The Growth of Logical Thinking. New York: Basic Books, 1968.

win, K., "Psychology of Success and Failure," Occupations, XIV (1936), pp. 926-930.

.rstein, B., "The Relationship of Grade Expectations and Grades Believed to be Deserved to Actual Grades Received," Journal of Experimental Education, XXXIII (1965), pp. 357-362.

nnington, L. A., "Shifts in Aspiration Level After Success and Failure in the College Classroom, "Journal of General Psychology, XXXIII (1940), pp. 305-313.

.ckup, A. J., and W. S. Anthony, "Teachers' Marks and Pupil's Expectations: The Short-term Effects of Discrepancies Upon Classroom Performance in Secondary Schools, "British Journal of Educational Psychology, XXXVIII (1968), pp. 302-309.

senfeld, H. and A. Zander, "The Influence of Teachers on Aspirations of Students," Journal of Educational Psychology, LII (1961), pp. 1-11.

tter, J. B., "Level of Aspiration as a Method of Studying Personality, I. A Critical Review of Methodology," Psychological Review, XLIX (1942), pp. 463-474.

mner, F. C., and E. C. Johnson, "Sex Differences in Levels of Aspiration and in Self-estimates of Performance in a Classroom Situation," Journal of Psychology, XXVII (1949), pp. 483-490.

lfe, R. B., "Perceived Locus of Control and Prediction of Own Academic Performance," Journal of Consulting and Counseling Psychology, (In press).